From: AAAI Technical Report SS-03-08. Compilation copyright © 2003, AAAI (www.aaai.org). All rights reserved.

Linking Video and Text via Representations of Narrative

Andrew Salway, Mike Graham, Eleftheria Tomadaki and Yan Xu

Department of Computing, University of Surrey
Guildford, Surrey, GU2 7XH
United Kingdom

a.salway@surrey.ac.uk

Abstract

The ongoing TIWO project is investigating the synthesis of language technologies, like information extraction and corpus-based text analysis, video data modeling and knowledge representation. The aim is to develop a computational account of how video and text can be integrated by representations of narrative in multimedia systems. The multimedia domain is that of film and audio description – an emerging text type that is produced specifically to be informative about the events and objects depicted in film. We suggest that narrative is an important concept for intelligent multimedia knowledge management. We then give an overview of audio description for film and discuss the integration of video and text data in this context.

Introduction[1]

Story understanding systems have focused typically on relatively short and simple written texts, like newspaper stories and children's stories. However, more elaborate storytelling can occur in various combinations of media and it is interesting to consider how the same story can be conveyed and understood in, for example, a moving image and in spoken or written language. Consideration of the relationship between different kinds of media is also required for the generation of enriched multimedia information sources. A computational understanding of the link between visual and textual information is necessary so that media processing technologies can be integrated to extract and combine rich representations of media content at a high level. Such representations facilitate more intuitive and personalized access and navigation of media artifacts for users, and facilitate further knowledge management services such as knowledge mining.

The need to integrate media representations, at different levels of abstraction, leads to the idea of a *media interlingua* (Maybury 1997). For example, low-level integration may involve representations of pixel distributions, sound wave features and keyword frequencies that can be expressed numerically. Higher level 'semantic media content' is generally equated with named events, objects, their attributes and their spatio-temporal relationships for which various logics have been proposed in the multimedia literature. However there are higher levels still that are not yet dealt with in multimedia systems. For example there are potential applications for which it may be necessary to represent sequences of connected events, causal relationships between events and the intentions of the people involved. Whilst media such as films are an obvious example, this may also be relevant for other kinds of systems, for example tracking business news. We suggest that theories about *narrative*, alongside existing knowledge representation schemes, can guide the development of media independent representations to integrate media at this level.

Extracting semantic content from still and moving images directly is not feasible in general cases, however successful approaches have been developed that extract information from text accompanying visual information, such as speech in video, closed captions and scripts. The *Informedia* system indexes intervals in news broadcasts and documentary programs by keywords that are extracted from speech and closed captions (Wactlar et al. 1999). Textual information from TV sit-com scripts was combined with visual features, through user interaction, in order for a system to locate scenes containing a particular character (Wachman and Picard 2001).

Recently another kind of text has begun to accompany an increasing number of television programs and films. Audio description is provided for the visually impaired with some television programs and with films in some cinemas and on VHS/DVD releases. In between existing dialogue a describer gives important information about on-screen scenes and events, and about characters' actions and appearances. In effect the story told by the moving image is retold in words. Audio description provides a good research scenario for exploring the integration of multiple media streams in terms of narrative. This time-coded text is produced specifically to be informative about the important objects, characters and events depicted by the moving image.

This paper considers how video data, specifically films, and text data, specifically audio description, may be

[1] Copyright 2003, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

integrated via representations of narrative, and to what extent narrative information can be extracted automatically from audio description.

Narrative and Intelligent Multimedia Knowledge Management

Narrative is a multi-faceted phenomenon studied by philosophers, literature and film scholars, psychologists, linguists, cognitive scientists and computer scientists. The study of narrative is concerned with how narrative forms of media are created, how different kinds of media can convey narratives, and, how narratives are understood. For some researchers, narrative abilities considered both as a mode of thought and of discourse are fundamental to intelligence. It has been suggested that "we organize our experience and our memory of human happenings mainly in the form of narrative", so it follows that, regarding the issue of how to represent 'reality', it may not be sufficient to "equate representations with images, with propositions, with lexical networks, or even […] sentences" (Bruner 1991:4-5). Within computer science the idea of narrative has been considered important in the design of story understanding and generation systems, human-computer interaction, virtual worlds for entertainment, education, and, we would like to argue it is important for intelligent multimedia knowledge management.

Narrative can be defined as "a chain of events in cause-effect relationships occurring in time and space" (Bordwell and Thompson 1997:90). In dramatic works the agents of cause-effect are the characters and their goals, beliefs and emotions. The events that comprise a narrative are not just those explicitly presented in some piece of media, but also events that are inferred by its audience. A distinction is made between, *plot* – comprising the events depicted, and *story* – a 'deeper' structure including events that can be assumed and inferred. Thus a story can be told using different plots, and its follows that the same story can be told using different media, or combinations of media.

For example a story told in a novel can be turned into a film. The novel may give more direct descriptions of a character's cognitive state but a film audience must infer them from the character's behavior, facial expressions, body language and dialogue. Such inferences are based on a combination of previous events in the story, general world knowledge and knowledge of stereotypical situations for that film genre. The film audience is constantly revising their understanding of the film as new events are presented, previous information is recalled and expectations for future events are changed. Similar processes may be going on as someone follows a developing news story over several days. Understanding a story about a company take-over attempt requires reasoning about information presented by multiple media streams (newspapers, television, radio), knowledge about previous events and the intentions of the organizations, and reference to similar situations in the past.

Perhaps then intelligent multimedia knowledge management systems would benefit from some narrative capabilities. Consider a video player that could visualize a film's storyline and explain why a character behaved the way they did: or, a news browser that could retrieve similar stories and their outcomes, not by keywords, but by patterns of events and intentions. In computational terms narrative comprehension requires the extraction and amalgamation of knowledge about characters, objects and events from multiple media streams – for film these include the moving image, the dialogue / sub-titles, the screenplay / script and now, audio description. Further processing may be supported by existing knowledge representation schemes for temporal, spatial and causal reasoning; including reasoning about intentions, emotions and beliefs.

Audio Description

Audio description enhances the enjoyment of most kinds of films and television programs for visually impaired viewers, including dramas, situation comedies, soap operas, documentaries, and sometimes live news broadcasts. Furthermore, there is considerable potential for audio description to be used by the whole television and film audience, for example to 'watch' a film on an audio CD or on WAP devices with little or no visual display. In the gaps between existing speech, audio description gives key information about scenes, events, people's appearances, actions, gestures, expressions and body language so that in effect the story conveyed by the moving image is retold in words.

Audio description is normally scripted before it is recorded. An audio description script is thus a text that is 'written to be spoken' and includes time-codes to indicate when each utterance is to be spoken. The following is an excerpt from the audio description script for *The English Patient*:

[11:43] Hanna passes Jan some banknotes

[11:55] Laughing, Jan falls back into her seat

[12:01] An explosion on the road ahead

[12:08] The jeep has hit a mine

Describers take care to ensure that audio description interacts with the existing dialogue and sound effects. They are also careful to strike the right balance between frustrating the audience with insufficient information to follow the story, and patronizing them by spelling out obvious inferences. Audio description is produced to be a surrogate for the visual component of television and film material and as such is different in character from scripts and screenplays, and from radio dramas and talking books. The linguistic features of audio description vary between program types and film genres, due in part to the time available for description, and in part to the expected audience; for example a describer will give simpler and more extensive description for a children's program.

Legislation and regulation mean that audio description is becoming increasingly available in countries like the UK, US, Canada, Germany and Japan. In the UK, audio description is prepared by trained professionals who follow established guidelines. Software is available to assist in the preparation of audio description scripts, the recording of audio description and its synchronization with program and film material. Audio description is provided via digital television broadcasts, in some cinemas and on some VHS and DVD releases of films.

To produce an 8000-word audio description for a 2-hour film may take 60 person hours, with many viewings and more than one describer: however a 30 minute soap opera which is almost full of dialogue and has familiar scenes and characters may take only 90 minutes to describe. Current computer systems to assist the production of audio description present video data on screen alongside a window that marks time-coded speech-free intervals into which the describer can type their descriptions. Once the script has been completed and reviewed it is spoken, recorded and synchronized with the video data by time-codes.

Guidelines for audio description suggest the use of the present tense and simple sentences, and the avoidance of ambiguous pronominal references. This restricted language, the presence of time-codes and the relatively straightforward chronological order of the texts make audio description scripts a good starting point for extracting information about the characters and events depicted in the moving image. As part of our research we are collecting audio description scripts: the Surrey Audio Description Corpus currently includes British English audio description scripts for 59 films provided by three different organizations. Linguistic variance and diversity are major issues in sampling for corpus design. The corpus is designed to be representative of 9 film genres, Table 1. Two audio description experts determined these genres, based on their opinions of how the language used for audio description might vary.

Film Genre	**Number of audio description scripts**
Action	9
Children's animation	7
Children's live action	3
Comedy	6
Dark	9
Miscellaneous	8
Period drama	6
Romantic	6
Thriller	5
TOTAL	59

Table 1: The Surrey Audio Description Corpus currently includes audio description scripts for 59 films spread across 9 genres.

Audio description refers to states of affairs that are being depicted on-screen at, or near to, the moment it is spoken; remember that the describer must work around existing dialogue. The following utterances exemplify the kinds of things described for films including a scene, a character's introduction, a character's physical features and clothing, and an action; time-codes have been removed because the utterances are not contiguous.

Beneath the aircraft the evening sun throws deep shadows amongst the soft rolling sand dunes.

A young French-Canadian nurse, Hana, adjusts the belt of her uniform.

She is wearing a simple white dress and her blonde hair is drawn back from her pale face.

She struggles with a soldier who grabs hold of her firmly.

The language of audio description is rich in information concerning the characters and their external appearance, but information about their cognitive states is not described directly. However some insight into a character's cognitive state is given when it is being conveyed visually and hence described, as by the adverbs in the following utterances.

She is resting peacefully on her side, as if asleep, one arm curled beside her face.

She peers anxiously through the window.

Another characteristic of audio description is the announcement of a new scene with temporal and / or spatial information, for example 'Italy, 1944', 'Later, the patient sings to himself'. The expression of temporal information is discussed further in the next section.

Integrating Film and Audio Description

TIWO (Television in Words, 2002-5) is an ongoing project investigating the synthesis of language technologies, like information extraction and corpus-based text analysis, video data modeling and knowledge representation to develop a computational account of narrative in multimedia systems. As part of this project, the AuDesc system is being developed to assist in the preparation of audio description, and to customize audio description for different audiences. More generally we are interested in extracting information from audio description to generate representations of video content for retrieval and browsing, and to visualize and identify patterns of characters' emotions that are crucial to stories.

Retrieving and Browsing Film Clips

AuDesc will extend the object-oriented KAB system (Knowledge-rich Annotation and Browsing) that processes already time-coded text alongside a digital video library in order to attach machine-executable annotations to video data (Salway and Ahmad 1998). KAB was used to access a digital library of dance videos, for which spoken commentaries were elicited from dance experts, and subsequently transcribed and time-coded; note that though similar in intent, these commentaries are not audio description per se. The user's view of moving images and texts in KAB is through a graphical user interface that combines annotation with retrieval and browsing functions. Keywords, from semi-automatically compiled term lists, are located in time-coded text fragments in order to label intervals of video data automatically. The term lists are compiled by analyzing a corpus of texts to identify terms that appear unusually frequently, for more details of this method see (Ahmad and Rogers 2001). Users can make keyword queries and can navigate the moving image by highlighting fragments of the text.

The AuDesc system, Figure 1, currently works in the same way as KAB, i.e. annotations comprise a start-time, an end-time and a key-word: this kind of annotation does not require a fixed segmentation of the video stream and allows for the layering of annotations to capture (some of) the multiplicity of meanings in a moving image. An occurrence of a term in a manually time-coded text fragment triggers the generation of an annotation with start and end times a fixed number of seconds either side of the time of the text fragment. Keyword annotations can be generated from lists of events or alternative references to the film's main characters. The integration of film and audio description is manifested as links between video intervals and time-coded text fragments. To develop the AuDesc system further we are interested in: (i) dealing with temporal information in audio description; and, (ii) implementing a 'narrative-like' intermediate representation to link video and text data. Our ideas about these issues are discussed in the following two sub-sections.

Temporal Information. A general task in the integration of video and text data is to ascertain the video intervals that text fragments refer to, and the temporal relationships between the events depicted in the video, and described in the text. In order to integrate audio description text with film at a semantic level it is necessary to deal with film in terms of the shots and scenes by which it is structured. It is also important to recognize two timelines: film time, i.e. the time it takes to watch the film; and, story time, i.e. the time in which the events depicted take place. Figure 2 shows how a film can be modeled in terms of shots that are defined as continuous pieces of filming, and scenes that are characterized by each having a unique combination of location and time. The story timeline is shown in parallel with layers of events taking place. Of course the relative position of events may differ between the two timelines, e.g. the film may depict events in a different order than they happen in the story, and events that are happening at the same time but in different locations will be depicted in different scenes. For video retrieval purposes it is important to maintain temporal relationships between events; different sub-sets of Allen's relationships have been applied in the video database literature (Allen 1983).

The structure of film provides some useful constraints when dealing with temporal information. It is reasonable to assume that all events depicted within a scene take place close together in the story timeline, and are likely to form larger events (information about scene boundaries and shot changes may be available from sources like film scripts and automatic video analysis). When considering how events are depicted at the shot level it is important to note film-making techniques that are used to convey that an event is taking place, or has taken place, without showing it in its entirety; a director may choose to portray only the end result of an event and allow the viewer to infer that the event took place.

Figure 1: Video data and audio description integrated in the current implementation of AuDesc. The user can search for video intervals with a keyword search. The user can also browse the film by highlighting text fragments.

The text is shown in Figure 2 as a series of time points that indicate the time at which the speaker starts the utterance (assuming a temporally aligned text, like an audio description). We have previously specified three tasks for the extraction of temporal information from audio description so that the audio description can be more closely integrated with film video data (Salway and Tomadaki 2002). The three tasks are to associate utterances with video intervals, to extract relationships between events and to establish the time at which scene is set.

It is necessary to associate an utterance with the video interval for which it is true, be it a shot, scene or some other interval. Given a time-coded text fragment it is relatively straightforward to associate it approximately with a video interval. In some cases it may be necessary to be more precise about at least one of: start time, end time or duration. Consider, for example, a system attempting to learn associations between visual features in a video stream and descriptions in accompanying text.

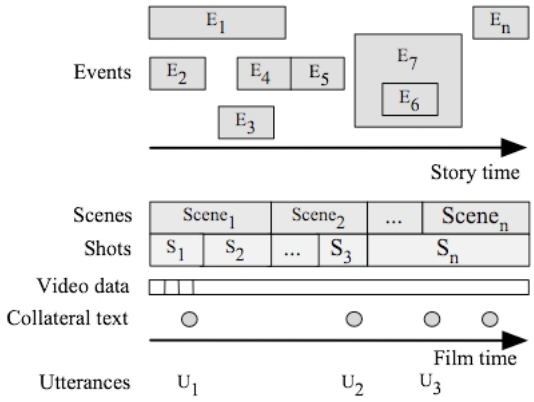

Figure 2. The organization of a film's content in terms of shots and scenes in film time, and the events that comprise the semantic video content in story time. Text such as audio description is aligned with the video data in film time.

It may help to consider some of the aspectual features of events (Comrie 1976). Whether an event has internal structure (*punctual* / *durative*), gives some indication of its duration, e.g. 'jump over' versus 'running'. Knowing about an event's end result, if it has one, (*telic* / *atelic*) gives information about its completion, e.g. 'Tom eats the cake' implies that the 'baking' event is complete. Analysis of the Surrey Audio Description Corpus indicated that the verbs 'start, stop, begin, finish' occur between 5 and 65 times more frequently in audio description than they do in general language (as exemplified by the British National Corpus): as such these may be useful cues for information about event boundaries. The adverb 'still' is also frequent and indicates an event is ongoing.

The second task is to extract temporal relationships between events in film time; here we only consider relationships holding within a scene. In audio description events are normally mentioned in the order they are seen on-screen, but events may occur simultaneously, or there may be stylistic reasons to mention them out of order. The conjunction 'as' was used widely in the corpus to indicate events happening at the same time – 'the children play as the crowd moves away', and it sometimes hints at a causal relationships – 'she continues to hide as the monster approaches'. The words 'then' and 'now' would seem to be redundant in as much the nature of audio description implies that events are mentioned in order and at time of occurrence. However they occurred frequently in the corpus and seem to imply more than basic temporal information. For example, 'Sarah chops the tomatoes, then fries an egg' implies that the end of the first event meets with the start of the second, and 'Jane is dancing with George, now she is dancing with her cousin'

highlights a contrast between the two events.

The third task is to establish the time at which a scene is set in story time. However our analysis found relatively few mentions of non-specific times of day, '*night, morning, evening, dusk, dawn*', and non-specific times of year such as months, seasons and festival days. There was almost no mention of specific times and dates. Perhaps it is the case that for many films the viewer need only understand a very general time period and this may be conveyed by costumes, props, and for times of day, lighting – all of which would be mentioned in audio description.

Having established a picture of the kinds of temporal information expressed in audio description we are currently specifying an annotation scheme for an information extraction system. The scheme will need to annotate temporal, and later other event-related information, in audio description. The TimeML annotation scheme being developed for the TERQAS project () provides a useful starting point but some extensions will be required. For audio description there is a need to maintain two timelines, record aspectual information regarding the internal structure of events and end-states, and sub-event and causal relationships.

Intermediate Representations to Link Video and Text. Recently various researchers have proposed the use of knowledge representation schemes in multimedia systems, including the use of conceptual graphs to link video data and text data, the use of semantic networks to browse between related events in a video, and the use of 'plot units' to create video summaries interactively. Here we consider some of these proposals with a view to developing a representation to integrate film and audio description at a 'narrative' level.

The *Smart VideoText* video data model uses conceptual graphs to capture semantic associations between the concepts described in text annotations of video data (Kokkoras et al. 2002). The representations form a knowledge-base about the named entities described, say in the speech / closed-captions: in one example three conceptual graphs are shown representing knowledge about a skyscraper and its designer. The aim is to facilitate query expansion and video browsing based on the relationships between entities: note that the emphasis seems to be on entities, rather than events, and it appears that the relationships hold in 'reality', rather than in the 'story world' of the video. This makes the model well suited to news and documentary programs, but perhaps less appropriate to film.

Semantic networks, with spreading activation, have been used to represent the causal dependencies between actions and events depicted in a movie (Roth 1999). One example network includes representation of a character pressing a detonator, and this event then causing an explosion. The user of the system can browse between related events when watching the video, for example to see why something happened; some general knowledge in the form of 'is-a' relationships is also stored. The focus of Roth's work was on the main entities visible in a video, their actions and attributes: this does not address characters' cognitive states and means that reasoning about why there was an explosion, for example, cannot go beyond the fact that someone pressed a detonator.

A scheme proposed by Lehnert (1981) as a technique of memory representation for story summarization has subsequently been applied to automatic text summarization systems, and more recently it was extended to include story threads in a system for browsing and interactively summarizing films (Allen and Acheson 2000). Compared with the semantic networks used by Roth, this representation of plot units is more character-oriented. It captures a character's cognitive states expressed as positive and negative emotions about sequences of actual and potential events. Depending upon their individual intentions, characters have the same or opposite feelings about an event.

Consider a sequence of events, from *The English Patient*, in which a pilot is shot down by enemy gunners and his injuries include memory loss – he is later found and cared for by nomads, and later in a military hospital he is questioned by an officer. These events are depicted in three non-contiguous scenes: the audio description for the first scene is as follows.

> [4:00] Clearly visible against the cloudless sky the plane flies on over the rolling sand hills.
>
> [4:04] German gunners spot the aircraft.
>
> [4:15] Bullets tear holes in the fuselage.
>
> [4:22] The plane catches fire.
>
> [4:26] In the rear cockpit, the pilot is enveloped in flames.
>
> [4:30] His clothes on fire, he struggles desperately to escape from the burning aircraft.

Using a semantic network, following Roth, it is possible to manually capture, among other things, the relationship between the shooting and the injury, Figure 3a. Of course the relationship 'CAUSE' is a controversial one, and is applied more intuitively to some pairs of events than others: for example, whilst being injured and being rescued are related, perhaps it will be necessary to refine the representation of their relationship.

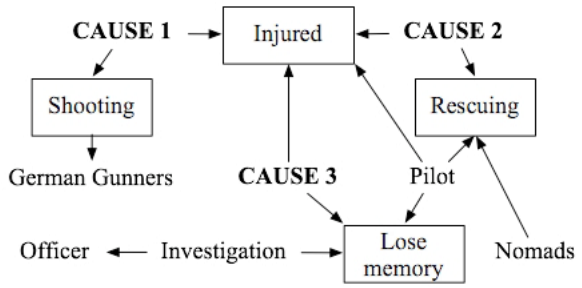

Figure 3a: A sketch of a semantic network, following (Roth 1999), for major events in three scenes of a film.

Following Lehnert's scheme it is possible to also capture the fact that the gunners wanted to harm the pilot, and that the nomads wanted him to get better, Figure 3b. The elements of the scheme are: M (mental state), + (positive event), - (negative event), m (motivation link), a (actualization link), t (termination link) and e (equivalence link). These elements can be combined into 15 primitive plot units that characterize stereotypical situations such as 'Problem', 'Loss' and 'Resolution'. The gunners view the pilot's flying negatively, and this motivates them into an action (shooting) which they view positively and the pilot views negatively, and he tries but fails to escape from his plane: this leads to a state in which the pilot is injured and the gunners' initial problem is resolved.

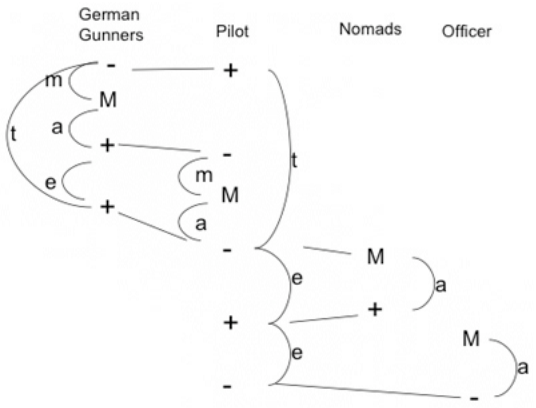

Figure 3b: A representation for the same three scenes following Lehnert (1981).

We are currently prototyping a video browsing system in which to evaluate semantic networks and plot units for representing films, and investigating the extent to which their instantiation may be automated through the processing of audio description.

Identifying Patterns of Emotion

A complete representation of a film should contain characters' cognitive states and the development of the story, as well as the physical objects and events depicted on-screen. Audio description provides an opportunity to investigate representations of the cognitive state of characters and story development in film, with the prospect of facilitating novel forms of querying, browsing and summarization. Our initial work is focusing on characters' emotional states. The aim is to visualize and subsequently classify dramatically important sequences and stereotypical emotional patterns in order to index video data at this level of semantic content. The approach is to analyze temporal patterns of key 'emotion words' in audio description.

Drama is rich in emotion – for a drama to engage an audience emotional states and behaviors must be displayed by characters in a consistent and overt fashion. In keeping with the tradition of filmed drama to show rather than to tell, it is rare that a character will make an explicit and honest declaration of his cognitive state. However, if the character shows an emotion, say a startled reaction or a look of contentment, this manifests quite directly in audio description. Emotional states are indicated with adjectives like *nervous*, adverbs like *desperately* and *nervously*, and nouns like *love* and *relief*. The order and frequency of such 'emotion words' in an audio description can perhaps be analyzed to extract information about patterns of emotion.

The cognitive theory of emotion due to Ortony, Clore and Collins (1988) provides a classification of emotion types that are cognitively linked to beliefs, goals and behavior. Emotions are treated as positive and negative evaluations of events, agents and objects with respect to important goals of the self. This provides a theoretically-grounded link that can facilitate between emotions and other cognitive states. For example, FEAR arises when a person believes a prospective event is likely to occur that will have a negative effect on an important goal. The emotion type RELIEF arises when this event fails to materialize. Example tokens are listed for the 22 emotion types in the theory; a sample is given in Table 2.

Charting the occurrence in audio description of tokens from the complete lexicon provided by Ortony and colleagues appears to give a visualisation of one aspect of a film's story. In the audio description for the film *Captain Corelli's Mandolin* there were 52 tokens of 8 emotion types, as shown in Figure 4.

Emotion type	Example tokens
FEAR	apprehension, anxious, dread, fear
HOPE	anticipation, expectancy, hope
JOY	contented, cheerful, delighted
RELIEF	relief
DISTRESS	distress, distraught, grief

Table 2: Some emotion types and tokens from Ortony, Clore and Collins (1988)

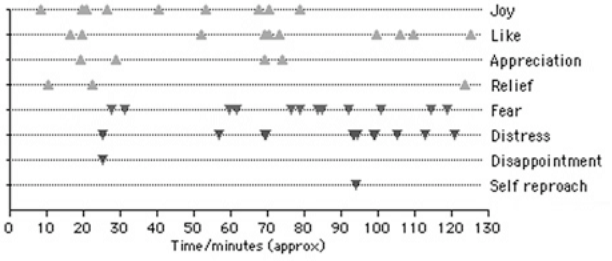

Figure 4: Occurrences of tokens for 8 emotion types in audio description for *Captain Corelli's Mandolin*

The story of this film concerns a love triangle between an Italian officer (Corelli), a Greek woman (Pelagia), and a Greek partisan (Madras) on the occupied Greek island of Cephallonia during World War II. A high density of positive emotion tokens appear between 15 and 20 minutes into the film, corresponding to Pelagia's betrothal to Madras. The negative emotion tokens which immediately follow are associated with the invasion of the island. The cluster of positive emotions between time 68 and 74 occur during scenes in which the growing relationship between Pelagia and Corelli becomes explicit. The group of FEAR, DISTRESS and SELF-REPROACH tokens between times 92 and 95 map to a scene in which German soldiers are disarming their former Italian allies, during which a number of Italians are gunned down. The clusters of emotion tokens appear to identify many of the dramatically important sequences in the film.

Further, there appears to be an intuitively appealing sequence of emotion types over the entire course of the film. For the first 75 minutes or so there is a clear sequence of JOY tokens, punctuated by a handful of DISTRESS and FEAR tokens. From this time onwards the JOY tokens disappear, and DISTRESS and FEAR tokens increase in frequency as violence descends on Cephallonia. Towards the end of the film, the last DISTRESS token appears, followed by a RELIEF token when Pelagia discovers her father has survived an earthquake, and a LIKE token as Pelagia and Corelli are reunited. There is some correspondence here with pre-theoretical notions of both story structure and conventional character behavior.

Encouraged by these observations, we are developing a method to visualizing and classifying stereotypical patterns of emotion types. This could support features like video retrieval based on story similarity. The method will involve the construction of a more comprehensive database of emotion tokens from general lexical resources. It would be desirable to denote emotional intensity in the database to distinguish between, say 'pleased' and 'ecstatic'. Techniques for time-series analysis may be applicable to classify patterns of emotion types.

Closing Remarks

Whilst a film may be perceived as the objects and events it depicts, it is more fully understood in terms of the story it tells, and so in order to facilitate more intuitive interaction with the user, a digital film system must be capable of creating and reasoning over representations of narrative. The emergence of film with audio description provides an opportunity to investigate representations of narrative. Film and audio description also provide an interesting scenario to investigate the integration of media processing technologies, like the language processing and knowledge representation technologies discussed in this paper, and potentially video and audio processing technologies too. Perhaps audio description could be used as 'training data' for a video processing system to learn associations between visual features and particular characters or events. Audio processing might provide further information about dramatically important sequences from the intonation of dialogue and from background music. There is also the need to extract information from dialogue to build a fuller representation of characters' behaviors: the fact that audio description is written around dialogue makes their interaction interesting.

Considering related work, the integration of vision and language has been studied in artificial intelligence as the "correspondence problem" of "how to correlate visual information with words" (Srihari 1995): importantly it is not just single words that are to be correlated with still and moving images (as, for example, in contemporary web search engines), but phrases, sentences and entire texts. There is also interest in how different modalities can combine in computer-based communication (McKeown et al. 1998). Looking to studies of human intelligence, the audio description task of putting images into words is reminiscent of investigations in cognitive psychology that used verbal reporting of images to

understand cognitive processes (Ericsson and Simon 1993) and studies of language production where subject groups gave verbal accounts of a film they had just seen (Chafe 1980). These studies suggest ways in which humans organize sequences of events and they help to explicate the linguistic realization of humans' narratives.

Acknowledgements

This research is supported by EPSRC GR/R67194/01. We thank the members of the TIWO Round Table (BBC, RNIB, ITFC and Softel) for sharing their knowledge of audio description, and the anonymous reviewer for comments on an earlier version of the paper.

References

Ahmad, K., and Rogers, M. 2001. Corpus Linguistics and Terminology Extraction. In Wright, S-E. and G. Budin, eds. 2001. *Handbook of Terminology Management*, Volume 2. John Benjamins: Amsterdam & Philadelphia.

Allen, J.F. 1983. Maintaining Knowledge About Temporal Intervals. *Communications of the ACM* 26 (11):832-843.

Allen, R. B.; and Acheson, J. 2000. Browsing the Structure of Multimedia Stories. *In Proceedings of the 5th ACM Conference on Digital libraries,* 11-18. New York.: ACM Press.

Bordwell, D., and Thompson, K. 1997. *Film Art: An Introduction*, 5th Edition. New York: McGraw-Hill.

Bruner, J. 1991. The Narrative Construction of Reality. *Critical Inquiry* 18:1-21.

Chafe, W. ed. 1980. *The Pear Stories: cognitive, cultural and linguistic aspects of narrative production.* Norwood, NJ.: Ablex.

Comrie, B. 1976. *Aspect: an introduction to the study of verbal aspect and related problems*. Cambridge University Press.

Ericsson, K. A., and Simon, H. A. 1993. *Protocol Analysis: verbal reports as data*. Cambridge, MA.: MIT Press.

Kokkoras, F.; Jiang, H.; Vlahavas, I.; Elmagarmid, A. K.; Houstic, E. N. and Aref, W. G. 2002. Smart VideoText: a video data model based on conceptual graphs. *Multimedia Systems* 8:328-338.

Lehnert, W. G. 1981. Plot Units and Narrative Summarization. *Cognitive Science* 4: 293-331.

Maybury, M. 1997. *Intelligent Multimedia Information Retrieval*. AAAI / MIT Press.

McKeown, K. R.; Steven K. Feiner, S. K.; Mukesh Dalal, M.; and Shih-Fu Chang, S. 1998. Generating Multimedia Briefings: Coordinating Language and Illustration. *Artificial Intelligence* 103: 95-116.

Ortony A., Clore G. L. and Collins A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Roth, V. 1999. Content-based retrieval from digital video. *Image and Vision Computing* 17:531-540.

Salway, A. and Tomadaki, E. 2002. Temporal Information in Collateral Texts for Indexing Movies. LREC Workshop - Annotation Standards for Temporal Information in Natural Language.

Salway, A.; and Ahmad, K. 1998. Talking Pictures: Indexing and Representing Video with Collateral Texts. In Proceedings of the Fourteenth Workshop on Language Technology - Language Technology for Multimedia Information Retrieval, 85-94.

Srihari, R. K. 1995 Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artificial Intelligence Review* 8(5-6):349-369.

Wachman, J. S.; and Picard, R. W. 2001. Tools for Browsing a TV Situation Comedy Based on Content Specific Attributes, *Multimedia Tools and Applications* 13 (3): 255-284.

Wactlar, H. D.; Christel, M. G.; Gong, Y.; and Hauptmann, A. G. 1999. Lessons Learned from Building a Terabyte Digital Video Library. *Computer* Feb, 1999:66-73.